DOCUMENT RESUME

ED 299 324 TM 012 359

AUTHOR Coffman, William E.
TITLE Perspectives on Intellectual Assessment--1986. Discussion.
PUB DATE 24 Aug 86
NOTE 7p.; Paper presented at the Annual Meeting of the American Psychological Association (Washington, DC, August 22-26, 1986).
PUB TYPE Speeches/Conference Papers (150) -- Reports - Evaluative/Feasibility (142) -- Information Analyses (070)

EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS Cognitive Ability; Cognitive Measurement; *Cognitive Tests; Elementary Secondary Education; *Intelligence Tests; *Research Reports; *Test Construction; Test Validity
IDENTIFIERS Criticism; K ABC Scales; *Stanford Binet Intelligence Scale; Wechsler Intelligence Scale for Children (Revised); *Woodcock Johnson Tests of Cognitive Ability

ABSTRACT

Three new tests presented as alternatives to the Wechsler tests for children were described in earlier papers of this session. This discussion reviews the earlier papers and comments on the developed tests: (1) the Woodcock-Johnson test; (2) the new Stanford-Binet; and (3) the K-ABC. The Woodcock-Johnson battery, state-of-the-art applications of traditional test construction methods, was evaluated only for its cognitive tests, without determining whether the total battery does its job. It has not been demonstrated that there is "treatment validity," that the cognitive measures provide specific guides to treatment. In fact, the same questions may be asked about the K-ABC, which attempts to measure abilities identified from studies of information processing and from neuropsychology and cognitive psychology, and the new Stanford-Binet test, a psychometric approach based on the hierarchical theory developed from factor analysis studies. For practitioners, the basic question concerns what the pay off is. Much research is needed to establish exactly what these tests measure and how patterns of scores relate to performance. (SLD)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED299324

DISCUSSION: PERSPECTIVES ON INTELLECTUAL ASSESSMENT - 1986

William E. Coffman
University of Iowa

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

☑ This document has been reproduced as received from the person or organization originating it.

☐ Minor changes have been made to improve reproduction quality.

• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY

William E. Coffman

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

TM012 359

American Psychological Association
Washington, D.C.
August 24, 1986

DISCUSSION: PERSPECTIVES ON INTELLECTUAL ASSESSMENT - 1986

William E. Coffman
University of Iowa

In summarizing his discussion of individualized tests of ability in the 3rd edition of Essentials of Psychological Testing (1970), Cronbach noted that the Wechsler tests were the dominant individual tests and anticipated some of the developments reported here today. Just how prophetic he was, or how far the developers of new tests have compromised in order to achieve continuity or to meet anticipated market demands will not be clear without more research and clinical experience, but it does seem that he was able to anticipate to a remarkable degree the message of Zimmerman's paper. She wrote

> The Wechsler's virtues are those of an aid to clinical observation rather than of a scientific measuring instrument. It is primitive in conception, embodying tasks that clinicians found useful in the early days of testing. Since these tasks were originally devised, psychological theory has advanced on many fronts; this theory did not enter into the design of the Wechsler and enters only indirectly into the profile organization and the interpretation. Measurement theory also has made advances by means of which a more efficient test could surely be engineered. It should be possible to obtain equally precise scores and equally rich observations in two-thirds the testing time. It should be possible to design a new structure for such a test with the aid of the "hierarchial" concepts of ability... (p. 252)

As Zimmerman says, "the real issue in the clinical interpretation of the WISC-R is that the years of experience with this test have established a standard which must be faced by each of the other tests on this panel." This is no idle claim. The Eighth Mental Measurements Yearbook notes 1585 references to the WISC and the Ninth adds an additional 299. Zimmerman recognizes that much of this accumulated "research" represents blind alleys or asks the wrong questions and that perhaps the most important value of this array of research "is to serve as an object lesson, so that newer tests may avoid the pitfalls of useless approaches which the

Wechsler has already explored so extensively. It would have been helpful if she had been more specific about what some of these pitfalls were. A close look at the other papers on today's panel suggests that it may not be an easy job to avoid some of these blind alleys in research.

It is probably fortunate that the three new tests seeking to compete with the WISC-R represent rather distinctively different approaches to the development of ability tests. The Woodcock-Johnson represents a "state of the art" application of traditional test construction methods, i.e. the authors assembled a large variety of homogeneous subtests that looked as if they might be useful and finally chose those that contributed significantly to the prediction of performances of interest. The new Stanford-Binet employs a psychometric approach based on the hierarchial theory developed from factor analytic studies. The K-ABC attempts to measure abilities identified from studies of information processing and from theory and research in neuro-psychology and cognitive psychology. As studies of the interrelationships among these measures of their relationships to other behaviors of interest to clinicians and educators accumulate, our insights into the nature of human abilities and of the significance of individual differences in abilities should grow significantly. In designing these studies, however, it will be very important to ask the right questions and to avoid repeating the errors of the past.

The Woodcock-Johnson battery was developed as a diagnostic tool for the educational setting, apparently for the purpose of identifying those pupils manifesting learning disabilities that might be treated if properly identified. The idea of a cognitive ability battery, an achievement battery, and an interest battery, all normed on the same population so that abnormal discrepancies could be identified represents a theory of some sort dispite the disclaimers of the authors. It is therefore too bad that the paper today deals only with the cognitive tests, i.e. with whether or not the cognitive battery measures are measures of ability or achievement, and not with the question of whether or not the total battery does the job it was designed to do. The demonstration that the Woodcock-Johnson is no more achievement oriented than the WISC-R seems beside the point. Nor is it particularly signficant to show that the cognitive battery contains some unique measures (Visual-Auditory Learning, Analysis-Synthesis, Concept Formation) or that these add to prediction in a step-wise multiple

regression. First, one needs to know that the achievement measures in the battery are ones that educators will accept as valid measures of what schools are trying to teach. Second, one needs to know whether the measures are making unique contributions, i.e. whether or not one might achieve essentially the same prediction with other combinations of subtests. Third, one needs to know whether whatever validities are established now will hold up over time. In this context, it might be a good idea to find out why tests of artificial language (similar in some ways to these new learning tests) were included in the experimental battery for the College Entrance Examination Board Scholastic Aptitude Test back in 1926 but never became part of the operational test. Finally, one needs to know whether the cognitive measures provide specific guides to treatments, to what one reviewer of the WISC-R refers to as "treatment validity."

> If one....defines a biased test as one which does not lead to effective educational interventions, then the WISC-R and most other standardized tests of intelligence and achievement are biased for all children, regardless of racial and/or ethnic group membership. In short, the WISC-R lacks treatment validity in that it does not enhance remedial interventions for children and for this reason should be replaced with assessment procedures which have greater treatment validity (Witt & Gresham, 9th MMYB, p. 1717).

If the Woodcock-Johnson can be shown to have "treatment validity" in this sense, the fact that it does not correspond to current models of psychological theory will be of little concern to educators.

To some extent, for the educational setting, the issue of treatment validity is important also for the other two batteries discussed today. They do appear to represent rather well two current psychological theories, and studies of the relationships among these measures and with other measures designed to reflect theoretical constructs will be of interest to psychologists studying such theories. For practitioners, however, the basic question is still "what is the pay-off?"

Thus, for practitioners, the fact that correlations among K-ABC scales and Luria-Nebraska scales are consistent with the underlying theory may simply mean that similar tests correlate higher with each other than with dissimilar tests. Rather, what they want to know is whether or not ability to answer such test questions tells them anything about how to go

about modifying the behavior of the child. Questions like the following are likely to be of particular interest:

1. Will the coordination between the K-ABC and the new adolescent and adult form allow one to monitor the intellectual development of children throughout the school years?

2. Do the sequential and simultaneous scales have treatment validity?

3. Will the planning scale in the new adolescent and adult form make a unique contribution to the understanding of abilities? Are there implications for how subjects in the secondary school should be taught?

4. What relationship will there be between the new measure of functional achievement and achievement in school subjects? What does this imply regarding how secondary school subjects should be taught -- or what subjects should be taught?

These question suggest perhaps that an explicit educational theory may be as important as a psychological theory in the design of future studies of these measures. The concept of learning disabilities implies that perhaps one of the most important functions of education is the development of abilities rather than simply the teaching of information content. If so, it becomes very critical to develop an understanding of how psychological theories are related to theories of teaching. For example, what are the implications if abilities now designated as <u>fluid</u> because they represent operations that have not been taught in school were to be practiced in a school setting until they become <u>crystalized</u>? Would such teaching result in greater intelligence or only in greater ability to produce correct answers to otherwise only <u>useful indicators</u> of ability?

The evidence reported by Scott suggests that the new Stanford-Binet, while dropping the age-scale format in favor of continuous subscales, may have been guided more by market estimates than by factor theory, with the theoretical formulation following development rather than preceding it. Thus, what is called the fluid/analytic subscore appears from the research data reported to be more a measure of spatial ability than of abstract/visual reasoning. Note that only two of the four subtests loaded on the abstract visual reasoning factor and that <u>matrices</u> failed to load on any of the four group factors. Note also that bead memory is distinguishable from the other three memory tests, suggesting that it might have been desirable to have produced two different memory scales.

The studies based on different samples seem to be ignoring the issue of treatment validity. The fact that intercorrelation patterns are the same from sample to sample seems much less informative than would have been studies showing diagnostic validity. The fact that the sample of 223 learning disabled subjects obtained a mean composite score that was 2.8 standard deviations below the mean of the normative sample doesn't say anything significant about the test. One would expect the same finding for any general test of ability. Of much more concern is the report that the standard deviation for the mentally retarded examinees was 16.2 and that the mentally retarded sample included 35% blacks while the gifted sample included only 6.1% black. These investigations seem to be asking the wrong questions. What one would really like to know is whether or not one can distinguish learning disabled from mentally retarded examinees....or whether there is anything in the patterns of subscores that implies anything about how to approach the teaching of individual children.

One final comment on Scott's paper. What is the justification for the fact that "conversion tables were developed so that an area score could be determined regardless of the number of tests that might be taken within each area" or that "similarly, a conversion table was developed so that the overall composite score could be determined based on one, two, three or all four area scores." Unless I'm missing something, this must mean that one can obtain an overall composite score by administering only one of the 15 subtests!

In summary, three tests that are competing with the WISC-R as measures of cognitive abilities have been described here. To some extent, what the tests are measuring may be inferred not only from the limited research generated during the process of developing the test and in the limited period since their publication but also from the accumulated research on predecessors of the several scales. On the other hand, much research will be needed in order to fill gaps in our understanding of what these tests are measuring and how the patterns of scores relate to performances of interest in applied settings. In designing such studies, it will be important to avoid following trails that have proved unproductive in the past and to focus on critical questions of "treatment validity" both in the educational and in the clinical setting.